Chaos

On the edges' PageRank and line graphs

Cite as: Chaos **28**, 075503 (2018); https://doi.org/10.1063/1.5020127
Submitted: 20 December 2017 . Accepted: 25 April 2018 . Published Online: 11 July 2018

Regino Criado, Santiago Moral, Ángel Pérez, and Miguel Romance

On the edges' PageRank and line graphs

Regino Criado,[1,2,3] Santiago Moral,[1,2,4] Ángel Pérez,[1] and Miguel Romance[1,2,3]
[1]*Department of Applied Mathematics, Rey Juan Carlos University, 28933 Madrid, Spain*
[2]*Data, Networks and Cybersecurity Research Institute, Rey Juan Carlos University, 28028 Madrid, Spain*
[3]*Center for Computational Simulation, Technical University of Madrid, 28223 Pozuelo de Alarcón, Madrid, Spain*
[4]*Cyber Security & Digital Trust, BBVA Group, 28050 Madrid, Spain*

(Received 20 December 2017; accepted 25 April 2018; published online 11 July 2018)

Two different approaches on a directed (and possibly weighted) network G are considered in order to define the PageRank of each edge of G with the focus on its applications. It is shown that both approaches are equivalent, even though it is clear that one approach has clear computational advantages over the other. The usefulness of this concept in the context of applications is illustrated by means of some examples within the area of cybersecurity and some simulations and examples within the scope of subway networks. *Published by AIP Publishing.* https://doi.org/10.1063/1.5020127

The impressive success with which complex networks have been used in different fields to model the interactions between the components of a variety of complex systems is well known.[1–25] Classic PageRank algorithm[26,27] constitutes a theoretical approach which originally provides a precise and quantitative measure of the relevance of a webpage by assigning authoritative weights to each page. Beyond its theoretical definition, this algorithm and its subsequent extensions have proven their effectiveness and versatility in a wide range of applications of all kinds where the arrangement and ranking of different elements is required.[27–42] In recent years, new tools and developments related to this concept and its applications have progressively emerged with the advances and new expansions of complex networks theory.[27–37] The aim of this paper is to compare the definition of edge's personalized PageRank of a network obtained via the classical PageRank algorithm with that obtained through the line graph associated with this network, showing that both approaches are equivalent. Several examples and simulations drawn from the field of cybersecurity and subway networks illustrate its usefulness in real applications.

I. INTRODUCTION

Line graphs have been studied for almost 90 years (the first time this concept appears in the literature was in 1932[43]). During this time several problems related to them have drawn people's attention[44–49] since, among other things, many properties and relationships in a network that depend only on the adjacent edge-to-edge relationship are immediately transferred to its associated line graph as equivalent properties that depend on the adjacent node relationship.

But surprisingly, when studying networks with a huge number of nodes and edges, line graphs have only been considered in a reduced number of studies and applications.[25,54,66–70] Particularly, in Ref. 54 it is shown that the study of centrality in complex networks within the context of urban design based on the primal graph representation gives us similar results than the corresponding analysis made using its associated line graph. In Ref. 25 different approaches to the definition of the line graph associated to a multiplex network were introduced and the potential utility of these approaches was illustrated by means of some simulations and examples in the context of subway networks.

The concept of line graph offers a good representation of the network properties when it is appropriate to give more importance to the edges of a network than to its nodes. Some examples of this comes from urbanism, transport networks and urban traffic networks.[25,50–54,61–65] But this concept also has a particularly significant application in the field of cybersecurity and intentional cyber-risk.[34] In fact, line graphs provide us with a natural way to define the concept of *accessibility* of a link within a computer network.[34] When computers are connected to each other and to the Internet, the security threats increase exponentially. In this regard, it is worth highlighting in a very significant way that in the early days of this discipline there were not many methodologies to assess risk. This is due, among other things, to the fact that, in order to determine the probability of a specific server, network or organization being attacked, it makes little sense to rely on observing how often this type of event has occurred in the past (its frequency in the past). Traditional methodologies determine the amount and criticality of vulnerabilities. The standpoint considered in Ref. 34 is based on firstly defining what is considered by potential attackers to be the most valuable assets, locate them inside a network, and then calculate what is the probability that attackers can choose to follow one path on the network instead of another, or jump to a particular server instead of another.

In this context, one of the main targets of this model is to determine what types of attacks are most likely to occur, leading to a model of cyber risk underpinned by the following pillars: game theory, based on Nash's analysis of equilibrium, and complex network theory, which determines the physical and logical structure where the game takes place.[34] In order to properly incorporate in the model the expected benefit of the attacker, three factors must be considered: the expected income (the cash equivalent value for the attacker), the estimated cost or effort that the attacker must make to achieve his objective, and the risk for the attacker along with its potential consequences. In this way, when focusing on the motivating elements for the attacker, it is necessary to consider three parameters in the model, called anonymity (easiness with which the identity of the attacker is determined), accessibility (easiness with which the attack can be carried out) and value (the potential profitability of the attack).

The accessibility of nodes and edges for this kind of network is calculated from the personalized PageRank algorithm[31,34] with a personalization vector suitable for modifying the PageRank vector.[31,36,42,71] But one of the problems that arises in this context is to relate the accessibility of the edges of our network with that of the nodes. In the following we will show that the accessibility of edges calculated from the application of the classical PageRank algorithm with a suitable personalization vector and the value obtained for edge's accessibility that follows from the use of the associated line graph $L(G)$ are equivalent. In this respect it is important to highlight the computational advantages obtained in this context when working with G instead of $L(G)$. Moreover, as we will see in Sec. VI, the use of a suitable personalization vector will allow us to study other types of problems related to the planning of the flow of passengers through the metro lines.

The structure of the paper is as follows: Section II is devoted to introducing the basic concepts and notations that will be used during the rest of the paper and to presenting the problem that will be addressed in the following sections, including an outline of the PageRank algorithm foundations. Section III describes the most natural way to obtain the ranking of the edges of a targeted network using the classic PageRank of nodes. In Sec. IV, after defining the weighted line graph associated to a directed and weighted line graph, this definition is used to introduce a PageRank approximation of edges different from the one previously considered. In Sec. V we prove that the two approaches contemplated are equivalent. Section VI shows the application of this result to the context of subway networks and Sec. VII is dedicated to presenting the conclusions of this work, emphasizing the computational advantages of one approach versus the other for applications.

II. PRELIMINARIES AND NOTATION

Throughout this paper we consider a *directed network* $G = (X, E)$, where $X = \{1, \ldots, n\}$ is the set of *vertices* or *nodes* and $E \subseteq X \times X$ is the set of *edges*. In the sequel we will also consider a directed and *weighted* network, i.e., a directed network $G = (X, E)$ joint to a function $w : E \longrightarrow [0, +\infty)$ in such a way that for each edge $(i, j) \in E$, the coefficient $w(i, j)$ is called *weight* of $(i, j) \in E$. If we have a directed network $G = (X, E)$ and this network does not have an associated weight-function, then we will say that G is a *non-weighted* network.

Given a directed and weighted network $G = (X, E)$ such that for each $(i, j) \in E$ its weight is given by $w(i, j)$, the (weighted) *adjacency matrix* of G is the matrix $A(G) = A = (a_{ij}) \in M_{n\times n}$ given by

$$a_{ij} = \begin{cases} w(i,j) & \text{if there exists an edge } (i,j) \in E, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

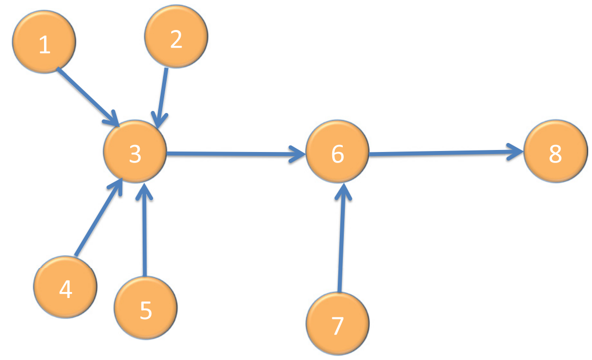

FIG. 1. Two edges have node 6 as destination but the frequency with which a random walker will pass through the edge $(3, 6)$ is much greater than that of the edge $(7, 6)$.

If $G = (X, E)$ is a non-weighted directed graph, its *adjacency matrix* is the matrix $A(G) = A = (a_{ij}) \in M_{n\times n}$ given by

$$a_{ij} = \begin{cases} 1 & \text{if there exists an edge } (i,j) \in E, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

i.e., we interpret each directed non-weighted network as a directed weighted network, by considering, for each $(i, j) \in E$, $w(i, j) = 1$. A more detailed explanation about this notation for directed and non-directed networks (weighted or not) may be found in Ref. 1.

Now, given a directed (and possibly weighted) network $G = (X, E)$, how can we define the PageRank of each edge of G?

It is possible to consider, *a priori*, two approaches to solve this problem:

1. By obtaining the PageRank of each edge from the PageRank of its nodes. In order to consider this approach it is needed to find a formula to give us for the real access frequency. In this respect it is not a good idea to consider the PageRank of an edge as the PageRank corresponding to its destination node (see Fig. 1).
2. By computing the PageRank (as usual) in a new auxiliary network (*line-graph* of the network) in which each edge of the original network be a node of this new auxiliary network. In this way, each edge of the original network is a node in the associated line-graph, and we will put an edge between two nodes of the line-graph if the destination and origin nodes of the corresponding edges match in the original network.

In the following sections, we will see how to compute the PageRank of each edge from the PageRank of its nodes, how to compute the edge's PageRank by using the associated line graph and finally we will demonstrate that both approaches are equivalent.

Note that this problem was already considered in Ref. 34 but that work did not provide a conclusive demonstration of the equivalence of both approaches.

In order to raise the problem that we are going to address properly, it is important to recall that the *classic PageRank algorithm* was originally employed by Brin and Page[26,27] to develop Google as a search engine to order the webpages, but in practice this web browser employs subtle modifications of this algorithm when ordering the webpages and, in fact, this algorithm has been extended to different contexts,[27,31,32,36,37] allowing to obtain new measures of centrality of complex networks based on this algorithm. Particularly important in our context is the idea of biasing the PageRank vector using a personalization vector, an idea already suggested in Ref. 27. Let us briefly recall underlying idea behind the foundations of this algorithm:

Random Walker hypothesis: If we move on the network in a *random* way, we will pass *more often* through the more *accessible* nodes. In order to mathematically model this idea, we must consider a specific type of Markov chains: *the random paths in a network*. In the context of applications, it is needed to consider an extension of the classic PageRank algorithm by considering a personalization vector and a weighted network. The idea is essentially as follows: a value $q \in (0,1)$ is fixed (traditionally in the case of Google $q = 0.85$). This value is the probability that a random walker does not change its trajectory jumping to other node of the network not connected with the previous one, instead of moving to a node connected directly by an edge with the current node. This value q is usually called *damping factor*. As we will see in Sec. VI this jump may be interpreted as the current random walker disappearing and a new random walker appearing in another place (another node) of the network. In the context of PageRank theory, this probability represents the case in which an imaginary surfer who is randomly clicking on links will eventually stop clicking. In addition a *personalization vector* $v \in \mathbb{R}^n$ is chosen, such that $v \geq 0$ (i.e. each coordinate of v is non-negative) and $\|v\|_1 = \sum_{i=1}^n v_i = 1$. The coordinate v_i of this vector represents the probability that the random walker, when jumping to a randomly chosen node with probability $1-q$ as described before, appears in the node i. In the The *classic PageRank* is obtained by taking the vector $v = \frac{1}{n}(1,\ldots,1)$, but for other applications it makes sense to consider other different personalization vectors. Once the path has started from any randomly chosen node in the network, for each $1 \leq i \leq n$, the *probability* to go *from* node i *to* node j is

$$\psi_{ij} = q\frac{a_{ij}}{\sum_k a_{ik}} + (1-q)v_j = qc_{ij} + (1-q)v_j. \tag{3}$$

In other words, we move to a neighbor directly connected by an edge with a probability q taking into account the weight of the corresponding edge, and we jump to a randomly chosen node with probability $(1-q)$, choosing the corresponding node j with probability v_j as is given in the personalization vector. Now, if we repeat the process indefinitely for $t = 1, t = 2, \ldots$, for each $t > 0$ we will get a vector $p_t = (p_t(1), \ldots, p_t(n))$, in such a way that each $p_t(j)$ give us the probability to be in the node j in the instant t. Therefore

$$p_t(j) = \sum_{i=1}^n p_{t-1}(i)\psi_{ij}. \tag{4}$$

So, if we consider previous expression in matrix form, with vectors p_t written as row vectors, we have that $p_t = p_{t-1}\Psi$, where $\Psi = (\psi_{ij})$ and if we navigate through the network in a random way, the frequency with which we pass through each node of the network is given by the vector $p \in \mathbb{R}^n$, where

$$p = \lim_{t\to\infty} p_t = \lim_{t\to\infty} p_0\Psi^t. \tag{5}$$

Without loss of generality, we will assume in the sequel that the matrix A has a non-zero coefficient in each row (that is, every node has an outgoing edge) and that every coordinate of the vector v is positive. These conditions let us ensure that the matrix Ψ is positive and *row-stochastic* or, which is the same thing, the sum of each row of the matrix Ψ equals 1. The existence of the previous limit is guaranteed by the fact that Ψ is positive and consequently, by using the Power Method, for each $0 \neq p_0 \in \mathbb{R}^n$ such that $p_0 \geq 0$, this limit exists and has the same value in all the cases.[74] In fact, this limit corresponds to the *unique* (except normalizations) positive eigenvector of Ψ (corresponding to the eigenvalue 1, since Ψ is row-stochastic).[74] Moreover, this vector is the one employed in ordering webpages, following the next definition:

Definition II.1. *If $G = (X,E)$ is a directed and weighted network with n nodes, $q \in (0,1)$ and $v = (v_1,\ldots,v_n) \in \mathbb{R}^n$ is such that $v > 0$ and $\|v\|_1 = 1$, then the* PageRank vector *of G with* damping factor *q and* personalization vector *v is the unique vector $PR(G,q,v) = PR \in \mathbb{R}^n$ such that*

(i) *$PR \geq 0$ and $\|PR\|_1 = 1$.*
(ii) *PR is an eigenvector corresponding to the eigenvalue is* 1 *of the matrix $\Psi = (\psi_{ij})$ given by*

$$\psi_{ij} = q\frac{a_{ij}}{\sum_k a_{ik}} + (1-q)v_j = qc_{ij} + (1-q)v_j, \tag{6}$$

i.e., $PR \cdot \Psi = PR$.

For each node $i \in X = \{1,\ldots n\}$ the PageRank *of the node i is the value $PR(G,q,v,i) = PR(i)$, the i*th *coordinate of the vector PR.*

Note that in the above definition each coordinate $PR(i)$ of the PageRank vector is interpreted as the frequency with which a random walker passes through the node i when it is randomly moving through the network, by taking q (at each step) as the probability to follow the network structure through the edges connected to the current node, and by taking the distribution given by the vector v if it jumps unexpectedly to another node of the network. It is important to consider random walks with positive jumping factor $q < 1$, since if $q = 1$ the matrix Ψ would be non-negative (instead of positive) and, as it is known, for guaranteeing the Power Method operates correctly it is necessary that the adjacency matrix of G be irreducible and primitive,[42,71] although in practice most of the real networks failure to satisfy this property.

III. EDGE'S PAGERANK VIA CLASSIC PAGERANK

Intuitively, if we have a directed and weighted network $G = (X,E)$, the frequency with which we use each edge $(i,j) \in E$ is related to the PageRank of the nodes i and j, since each time a random walker pass through the edge (i,j), this random walker also pass through the nodes i and j. In fact, to understand in depth the relationship between $PR(i)$, $PR(j)$ and the frequency of use of each edge (i,j), it is necessary to imagine the random walker as a random walker in a two-layered network[9–11] without random jumps (i.e., always following the structure of the network) as explained below.[36] If we have a mono-layer network $G = (X, E)$, the PageRank of this network with damping factor q and personalization vector v can be understood as the frequency in which we pass through the nodes of a two-layered multiplex network built as follows:

1. The top layer is a copy of the original network G.
2. The lower layer is a complete network in which all the nodes of G are connected between them in such a way that the weight of each edge (i, j) is $w(i, j) = v_j$. This layer is used to model the random jumps made by the walker when its movements does not follow the structure of the network.

In this biplex network we consider the following random walker:

1. In each step, we start by choosing the layer where the random walker is going to make the movement. With probability q the random walker will be in the top layer (i.e., the original structure of the layer G) and with probability $1 - q$ the random walker will be in the lower layer (*underworld*).

Now, if we fix a node $i \in X$ and $PR(i)$ is the PageRank in the network G, then it is immediate that

$$PR(i) = \sum_{j=1}^{n} qPR(i)c_{ij} + \sum_{j=1}^{n}[(1-q)v_j PR(i)]. \qquad (7)$$

What is really interesting in this expression is that the summands of the first sum compute the frequency with which the random walker pass through the edge (i, j) in the top layer (the layer corresponding to the structure of G), while the summands of the second sum show us the frequency with which the random walker pass through the edge (i, j) considered as an edge of the lower. Therefore, if we want to compute the frequency with which the random walker in the original network G pass through the edge $(i, j) \in E$ actually is

$$qPR(i)c_{ij}, \qquad (8)$$

and, therefore, normalizing so that the sum of all frequencies be equal to 1, considering all edges $(i, j) \in E$ only in the top layer, we have the following definition:

Definition III.1. *Given a directed and weighted network* $G = (X, E)$, $\alpha \in [0, 1]$ *and* $v \in R^n$, $v \geq 0$ *such that* $\|v\|_1 = 1$, *for each* $(i, j) \in E$ *we call PageRank of the edge* (i, j) *with jumping factor* q *and personalization vector* v *to the value*

$$PR[G, q, v, (i, j)] = PR(i, j) = c_{ij}PR(i). \qquad (9)$$

IV. EDGE'S PAGERANK VIA LINE GRAPH

In this section, in order to obtain the definition of the PageRank of edges in a natural way we consider a new auxiliary network in which the nodes are the edges of the original network with the appropriate weights: the *line-graph* $L(G)$ of a weighted and directed network. Line graphs have been widely studied in scientific literature[44,49,62,66,67] showing its usefulness in different contexts (see, for example, Refs. 25, 46, 47, 51, 52, and 55–61). To employ this mathematical tool in our context, it is necessary to define the line-graph $L(G)$ of a

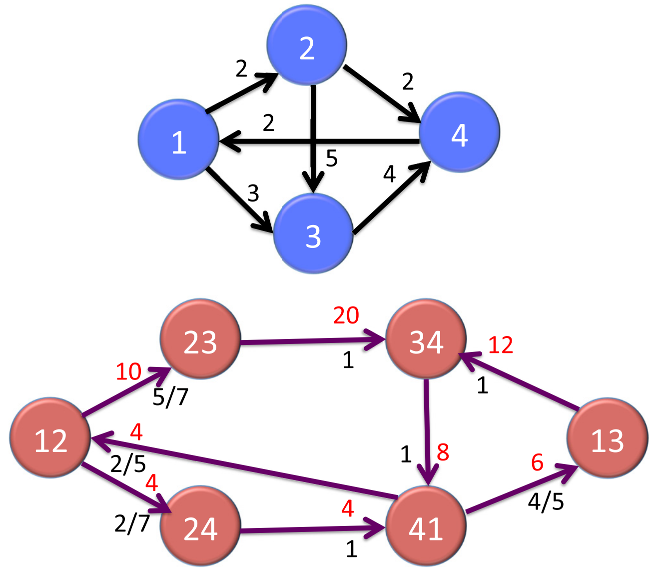

FIG. 2. Line graph of a weighted and directed network, including the weights and the normalized weights of the new edges.

directed and weighted network G. There are some alternatives in the literature employed to define this concept,[44,45,47,55,56] but we think that the following is the most appropriate alternative for the problem we are studying: since the weight of each edge is related to the frequency of use of that edge (this idea is inspired by the concept of accessibility in intentional complex networks[34,35]) and bearing in mind that each edge of $L(G)$ is identified with a path of length 2 in the original network G, the weight of each edge $[(i, j), (j, k)]$ of $L(G)$ will necessarily be related to the frequency of use of the path $(i, j) \to (j, k)$, i.e., it will be related to the product of the frequencies of use of the edges (i, j) and (j, k) (see Fig. 2). Following this idea we can give the next definition:

Definition IV.1. *If* $G = (X, E)$ *is a directed and weighted network,* the directed and weighted line-graph of G *is the network* $L(G) = (E, \tilde{E})$, *where*

$$\tilde{E} = \{[(i, j), (j, k)]; \quad (i, j), (j, k) \in E\} \qquad (10)$$

and where the weight for each edge $[(i, j), (j, k)] \in \tilde{E}$ *of* $L(G)$ *is given by the weighting function* $\tilde{w} : \tilde{E} \longrightarrow [0, +\infty)$ *whose expression is*

$$\tilde{w}[(i, j), (j, k)] = a_{ij}a_{jk}. \qquad (11)$$

It is important to highlight that if $G = (X, E)$ is a directed and non-weighted network, the concept of line-graph $L(G)$ as weighted network (with all the weights equal to 1) coincides with the classical concept we can find, for example, in Ref. 67. Now we have defined the concept of line-graph $L(G)$ of a directed and weighted network G, the concept of PageRank in $L(G)$ arises in a natural way and it makes perfect sense to consider the frequency with which a random walker passes through the corresponding edges, as it is shown by the following definition:

Definition IV.2. *If* $G = (X, E)$ *is a directed and weighted network with* n *nodes and* m *edges,* $q \in (0, 1)$ *and* $u = (u_1, \ldots, u_m) \in \mathbb{R}^m$ *such that* $u > 0$ *and* $\|u\|_1 = 1$, *then we call* PageRank vector *of the directed and weighted line-graph*

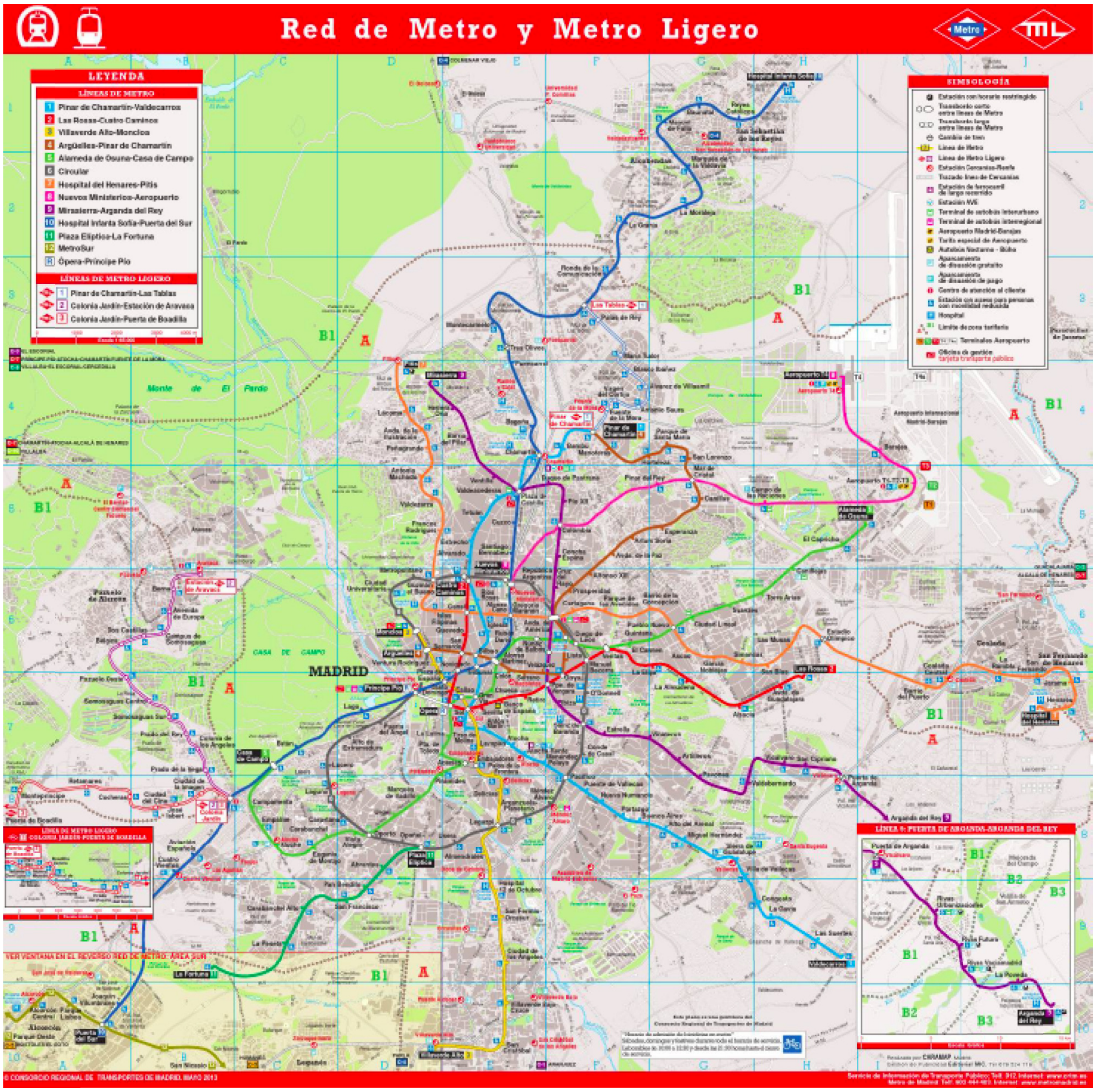

FIG. 3. Geographical map of the Madrid underground.[72]

$L(G)$ with damping factor *q and* personalization vector *u to the unique vector $LPR(G, q, u) = LPR \in \mathbb{R}^m$ such that*

- *$LPR \geq 0$ and $\|LPR\|_1 = 1$.*
- *LPR is an eigenvector (associated to the eigenvalue 1) of the matrix $\Phi = (\phi_{i\to j,k\to l})$ given by*

$$\phi_{i\to j,k\to l} = q\frac{b_{i\to j,k\to l}}{\sum_{(\alpha\to\beta)} b_{i\to j,\alpha\to\beta}} + (1-q)u_{k\to l}, \qquad (12)$$

i.e., $LPR \cdot \Phi = LPR$. Here $B = (b_{i\to j,k\to l})$ denotes the weighted adjacency matrix of $L(G)$.

For each node $(i, j) \in E$ the PageRank of the edge (i, j) is the value $LPR[G, q, u, (i, j)] = LPR(i, j)$ i.e., the (i, j)th coordinate of the vector LPR.

The following lemma will allow us to demonstrate the main theorem included in the following section and provides a more compact expression of the coefficients of the matrix Φ in terms of the coefficients c_{ij} from Sec. III.

Lemma IV.3. *Let Φ be the matrix as in Definition IV.2. For every pair of edges $i \to j, k \to l \in E$, we have the following equality:*

$$\phi_{i\to j,k\to l} = q\delta_{jk}c_{kl} + (1-q)u_{k\to l}, \qquad (13)$$

where δ_{jk} is the Kronecker delta.

Proof. Due to the way the line graph is constructed, we have that $b_{i\to j,k\to l} \neq 0$ only if $j = k$. Therefore if $j \neq k$ the result follows immediately.

Let us analyze the case $j = k$. We have

$$\frac{b_{i\to k,k\to l}}{\sum_\beta b_{i\to k,k\to\beta}} = \frac{a_{ik}\cdot a_{kl}}{\sum_\beta a_{ik}\cdot a_{k\beta}} = \frac{a_{kl}}{\sum_\beta a_{k\beta}} = c_{kl}.$$

From this equality the demonstration of the lemma is easily obtained. □

V. CLASSIC PAGERANK VS LINE-GRAPH'S PAGERANK

In this section, we show the relationship between the two values of the corresponding PageRank obtained for the same edge $(i,j) \in E$ in the two previous sections. In other words, we give a solution for the following problem:

Problem. *Given a directed and weighted network $G = (X, E)$ with n nodes and m edges, $q \in (0,1)$, and given the personalization vectors $v \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$, if we choose $(i,j) \in E$, what is the relationship between $PR[G, q, v, (i,j)]$ and $LPR[G, q, u, (i,j)]$?*

The answer to this problem is given by the following theorem:

Theorem V.1. *If $G = (X, E)$ is a directed and weighted network with n nodes and m edges, $q \in (0,1)$, $v \in \mathbb{R}^n$ is a personalization vector and $\tilde{u} \in \mathbb{R}^m$ is the personalization vector given by*

$$\tilde{u}_{i\to j} = v_i\, c_{ij}. \tag{14}$$

Then, for every edge (i,j), we have

$$PR[G, q, v, (i,j)] = LPR[G, q, \tilde{u}, (i,j)], \tag{15}$$

Proof. So, given a personalization vector $v \in \mathbb{R}^n$, let us define two auxiliary matrices $P \in M_{n\times m}$ and $S \in M_{m\times n}$ which will used throughout the proof. The rows of P will be indexed by nodes of G and the columns of P by edges of G. The opposite happens for S.

The coefficient $p_{j,k\to l}$ is defined as

$$p_{j,k\to l} = \begin{cases} c_{kl} & \text{if } j = k, \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

that is, $p_{j,k\to l} = \delta_{jk} c_{kl}$.

On the other hand, $s_{k\to l,j}$ is defined as

$$s_{k\to l,j} = \begin{cases} q + (1-q)v_j & \text{if } l = j, \\ (1-q)v_j & \text{otherwise.} \end{cases} \tag{17}$$

that is, $s_{k\to l,j} = q\delta_{lj} + (1-q)v_j$.

Now from P and S we construct a block matrix $T \in M_{(n+m)\times(n+m)}$ as follows:

$$T = \left(\begin{array}{c|c} 0 & P \\ \hline S & 0 \end{array}\right). \tag{18}$$

Next, we are going to show that

$$T^2 = \left(\begin{array}{c|c} 0 & P \\ \hline S & 0 \end{array}\right) \cdot \left(\begin{array}{c|c} 0 & P \\ \hline S & 0 \end{array}\right) = \left(\begin{array}{c|c} PS & 0 \\ \hline 0 & SP \end{array}\right) = \left(\begin{array}{c|c} \Psi & 0 \\ \hline 0 & \Phi \end{array}\right),$$

where Ψ is the matrix whose positive eigenvector is $PR = PR(G, q, v)$ (constructed like in Definition II.1) and Φ is the matrix whose positive eigenvector is $LPR = LPR(G, q, \tilde{u})$ (constructed like in Definition IV.2), with $\tilde{u}$ computed from v as in Eq. (14).

Let us start by showing that $PS = \Psi$. Consider the coefficient

$$\begin{aligned} (PS)_{ij} &= \sum_{k\to l} p_{i,k\to l} s_{k\to l,j} = \sum_{k\to l} \delta_{ik} c_{kl} s_{k\to l,j} \\ &= \sum_l c_{il} s_{i\to l,j} = \sum_l c_{il}[q\delta_{lj} + (1-q)v_j] \\ &= qc_{ij} + (1-q)v_j \sum_l c_{il} \\ &= qc_{ij} + (1-q)v_j = \psi_{ij}. \end{aligned}$$

On the other hand, let us show that $SP = \Phi$. Consider the coefficient

$$\begin{aligned} (SP)_{i\to j,k\to l} &= \sum_\beta s_{i\to j,\beta} p_{\beta k\to l} = \sum_\beta s_{i\to j,\beta} \delta_{\beta k} c_{kl} \\ &= s_{i\to j,k} c_{kl} = [q\delta_{jk} + (1-q)v_k]c_{kl} \\ &= q\delta_{jk} c_{kl} + (1-q)v_k c_{kl} \\ &= q\delta_{jk} c_{kl} + (1-q)\tilde{u}_{k\to l} = \phi_{i\to j,k\to l}, \end{aligned}$$

where last equality derives from Lemma IV.3.

Now consider the vector $y = (PR\,|\,LPR) \in \mathbb{R}^{n+m}$ which is, by the equality we have just proven for T^2, a positive eigenvector of this matrix associated to the eigenvalue 1. As the matrix T is row-stochastic, y is also an eigenvector of T associated to the eigenvalue 1, therefore, we have

$$\begin{aligned} (PR\,|\,LPR) &= (PR\,|\,LPR) \cdot T = (PR\,|\,LPR) \left(\begin{array}{c|c} 0 & P \\ \hline S & 0 \end{array}\right) \\ &= (LPR \cdot S\,|\,PR \cdot P) \end{aligned}$$

and in particular $LPR = PR \cdot P$. Now if we consider coefficient $i \to j$ we have

$$\begin{aligned} LPR(i\to j) &= (PR \cdot P)_{i\to j} = \sum_k PR(k) p_{k,i\to j} \\ &= \sum_k PR(k)\delta_{ki} c_{ij} = PR(i) c_{ij}, \end{aligned}$$

which is precisely the definition of $PR[G, q, v, (i,j)]$ which appears in Definition III.1. This concludes the proof. □

Remark V.2. *Last theorem shows that the two approaches studied in the two previous sections are essentially equivalent. More precisely, it allows to compute the PageRank in $L(G)$ (i.e., the edges' PageRank) from the PageRank in G (from the nodes' PageRank) and vise versa. Obviously this result allows us to calculate the PageRank of the edges of a directed and weighted network from the PageRank of its nodes, with the consequent computational advantages, as we will see in Sec. VI.*

Remark V.3. *Notice that, in our previous result, $\tilde{u}$ can be obtained from v by product with the matrix P which appears in the proof, specifically $\tilde{u} = v \cdot P$. We could considered the reciprocal problem to the one solved in the theorem, that is, given a personalization vector $u \in \mathbb{R}^m$, is it possible to find a personalization vector $\tilde{v} \in \mathbb{R}^n$ such that $PR[G, q, \tilde{v}, (i,j)] = LPR[G, q, u, (i,j)]$ for every edge?*

Let us consider the linear mapping $\pi : \mathbb{R}^n \to \mathbb{R}^m$ associated to the matrix P, that is $\pi(v) = v \cdot P$. Because of the

TABLE I. Ranking of the 15-top stations according to its PageRank during the morning traffic.

N.	Station	PR_i
1	Puerta del Sur	0.010539
2	Las Suertes	0.008506
3	La Poveda	0.008500
4	Reyes Católicos	0.008175
5	La Peseta	0.008088
6	Las Musas	0.008078
7	Arroyo Culebro	0.007996
8	Conservatorio	0.007996
9	Parque de los Estados	0.007996
10	Alonso de Mendoza	0.007994
11	Fuenlabrada Central	0.007994
12	Getafe Central	0.007992
13	Parque Europa	0.007992
14	Hospital de Fuenlabrada	0.007988
15	Juan de la Cierva	0.007988

TABLE II. Ranking of the 15-top stations according to its PageRank during the after-work traffic.

N.	Station	PR_i
1	Vodafone Sol	0.018731
2	Alonso Martinez	0.017056
3	Ópera	0.014599
4	Avenida de América	0.014358
5	Cuatro Caminos	0.012598
6	San Bernardo	0.012205
7	Príncipe Pío	0.011990
8	Diego de León	0.011968
9	Gran Vía	0.011834
10	Príncipe de Vergara	0.011823
11	Bilbao	0.011793
12	Arguelles	0.011751
13	Callao	0.011648
14	Goya	0.011644
15	Nuevos Ministerios	0.011497

properties of the matrix P, the map π is injective. If we start from a personalization vector $u \in \mathbb{R}^m$ which is in the image of π, i.e., it is a linear combination of the rows of P, then we can answer previous question in the positive. Just take $\tilde{v} \in \mathbb{R}^n$ as the unique vector such that $\pi(\tilde{v}) = w$. The equality between $PR[G, q, \tilde{v}, (i,j)]$ and $LPR[G, q, u, (i,j)]$ follows from the theorem in this section. Note also that $\tilde{v}$ can be easily computed from u in the following way:

$$\tilde{v}_i = \sum_k u_{i\to k}. \tag{19}$$

VI. APPLICATIONS FOR PREDICTING HUMAN MOBILITY IN SUBWAY NETWORKS

As an application of the results obtained in the previous section, we are going to analyze human mobility in the Madrid Metro System[72] (see Fig. 3) in order to locate the segments with the highest passenger flow on a standard working day, distinguishing between the morning and the afternoon time periods. In this regard, it is important to point out that the entire Madrid metro line covers a total of 301 stations with 26 interchanges, 13 lines, 294 km of network and through which 584 845 945 passengers circulated in 2016,[72] making it the ninth largest metro network in the world behind Shanghai, Beijing, London, Guangzhou, New York, Moscow, Seoul and Tokyo.[73] According to figures in the Metro de Madrid report, each of the more than 2.5 million daily passengers that Metro de Madrid has a 15-kilometre journey per day (counting the round trip), a journey in which they spend approximately 40 min on average (20 min and 7.5 km on each of the two routes).[72] This information is supplemented by the average distance between stations, which reaches a length of 108 485 m.[72] The average speed of the trains is 30 km per hour, which, in addition to placing it within the parameters of the most competitive meters in the world, allows us to estimate the damping factor for our personalized PageRank applied to this context. Taking into account the above data, the average number of stations traveled by a passenger is 11. So, in this context, we can deduce the damping factor corresponding to this situation, as follows:

$$11 = \mathbb{E}(\ell) = \sum_{k=0}^{\infty} k \cdot \mathbb{P}(\ell = k) = \sum_{k=1}^{\infty} k \cdot (1-q) \cdot q^k$$
$$= (1-q) \cdot q \sum_{k=1}^{\infty} k \cdot q^{k-1} = \frac{q}{1-q}.$$

So in this context the damping factor is $q = \frac{11}{12} = 0.91$.

As explained above, this value is the probability that a random walker does not change its trajectory jumping to other node of this network not connected with the previous one, instead of moving to a node connected directly by an edge with the current node. In our context, this jump may be interpreted as the current passenger disappearing (that is, the traveler gets off at that station) and a new passenger appearing in another node (another station) of the network.

To complete our model we are going to consider two personalization vectors, one for the morning and the other for the afternoon. In this regard, taking into account the geographical, economic and social aspects related to the service provided by the Madrid Metro System, we have divided all the metro stations into four types:

1. Stations located in urban areas with many residents, dormitory districts, surrounding suburbs, or nearby peripheral cities.
2. Stations located in the vicinity of the main work centers.
3. Transit stations.
4. Stations belonging simultaneously to the first two types.

In order to obtain the personalization vector for the morning, we assume that 80 percent of passengers enter the system through a type 1 and 4 station, and the remaining 20 percent enter through a type 2 and 3 station. On the other hand, in order to obtain the personalization vector for the afternoon, we assume that 80 percent of passengers enter the system through a station of types 2 and 4 and the remaining 20 percent enter through a station of types 1 and 3. The results obtained are reflected in Tables I–IV in which the corresponding PageRank

TABLE III. Ranking of the 15-top segments according to its PageRank during the morning traffic.

N.	Segment	$PR_{(i,j)}$
1	Valdecarros → Las Suertes	0.004590
2	Arganda del Rey → La Poveda	0.004587
3	Hospital Infanta Sofia → Reyes Católicos	0.004440
4	La Fortuna → La Peseta	0.004400
5	Estadio Olímpico → Las Musas	0.004395
6	Villaverde Alto → San Cristobal	0.004329
7	Las Suertes → La Gavia	0.004253
8	Las Suertes → Valdecarros	0.004253
9	La Poveda → Arganda del Rey	0.004250
10	La Poveda → Rivas Vaciamadrid	0.004250
11	Las Rosas → Avenida de Guadalajara	0.004235
12	Paco de Lucía → Mirasierra	0.004183
13	Pitis → Lacoma	0.004163
14	Reyes Católicos → Baunatal	0.004088
15	Reyes Católicos → Hospital Infanta Sofia	0.004088

TABLE IV. Ranking of the 15-top segments according to its PageRank during the after-work traffic.

N.	Segment	$PR_{(i,j)}$
1	Antón Martín → Atocha	0.003843
2	Antón Martín → Tirso de Molina	0.003843
3	Atocha → Antón Martín	0.003802
4	Atocha → Atocha Renfe	0.003802
5	Embajadores → Lavapiés	0.003785
6	Embajadores → Palos de la Frontera	0.003785
7	Palos de la Frontera → Delicias	0.003700
8	Palos de la Frontera → Embajadores	0.003700
9	Banco de España → Retiro	0.003652
10	Banco de España → Sevilla	0.003652
11	Tirso de Molina → Antón Martín	0.003631
12	Tirso de Molina → Vodafone Sol	0.003631
13	Lavapiés → Embajadores	0.003604
14	Lavapiés → Vodafone Sol	0.003604
15	Sevilla → Banco de España	0.003543

of the segments has been obtained, using the result of the previous section, from that corresponding to the stations. As can be seen, the rankings collected in these tables coincide with the geographical, sociological and economic aspects related to the service provided by the Madrid Metro Madrid System. So, for instance, the station listed first in Table I is "Puerta del Sur," which is the bottleneck of line 12, which is a circular line that runs through several dormitory municipalities in the south of Madrid, while "Las suertes" and "La poveda" are two of the main header metro stations on lines 1 and 9 (respectively) that cover important population centers. Moreover, the first stations in Table II ("Vodafone Sol," "Alonso Martinez," "Ópera,"...) correspond to important work areas located in the center of Madrid and surroundings, and the first segments of Tables III and IV correspond, respectively, to the metro segments in which there is a greater flow of passengers in morning and afternoon time, appearing in the top places, in the first case, segments close to important passenger residency areas and centers (Valdecarros → Las Suertes, Arganda del Rey → La Poveda,...), while in the second case, the respective segments (Antón Martín → Atocha, Antón Martín → Tirso de Molina,...) are located close to the most important workplaces.

VII. CONCLUSIONS AND FUTURE WORKS

From a definition of the concept of line graph of a directed and weighted network, we have demonstrated the equivalence between two different approaches to the personalized PageRank of a directed and weighted network. This result allows us to calculate the PageRank of the edges of a directed and weighted network from the PageRank of its nodes, with the consequent computational advantages, as it is easier to compute the personalized edges' PageRank from this result without the need to compute it on the corresponding line graph. The usefulness of this result in the area of cybersecurity and intentional cyber-risk is related to the computation of edge accessibility, one of the three basic parameters underpinning intentional risk.[34] Also, by means of some simulations on a real subway network, we show how to use a personalization vector suitable for biasing the PageRank with the aim of determining the segments with highest passenger flows in a subway network depending on the time zone under consideration.

A natural future task will be to study this equivalence for the different mathematical models for the line graph of a multiplex network[25] incorporating stations, sections and subway lines as differentiated elements into the study and simulations of passenger flows.

ACKNOWLEDGEMENTS

Authors gratefully acknowledge the Technological Risk Management Research Center (Centro para la Gestión Tecnológica del Riesgo, CIGTR) sponsored by the Rey Juan Carlos University and BBVA Group, and the I4S-URJC Chair on Information Security, Fraud Prevention and Technological Risk Management that encourage the base research on pain points of risk management within information systems.

This work has been partly supported by the Spanish MINECO under project MTM2014-59906-P, by the I4S-URJC Chair on Information Security, Fraud Prevention and Technological Risk Management and by the grant for the researching activity for excellence group GARECOM GI_EXCELENCIA 30VCPIGI11.

[1] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, Phys. Rep. **424**, 175–308 (2006).
[2] M. E. J. Newman, *Networks: An Introduction* (Oxford University Press, New York, 2010).
[3] E. Estrada, *Networks Science* (Springer, New York, 2010).
[4] M. E. J. Newman, SIAM Rev. **45**, 167–256 (2003).
[5] S. Wasserman, and K. Faust, *Social Network Analysis* (Cambridge University Press, Cambridge, 1994).
[6] R. Albert and A. L. Barabási, Rev. Mod. Phys. **74**, 47–97 (2002).
[7] S. H. Strogatz, Nature **410**, 268–276 (2001).
[8] Y. Bar-Yam, *Dynamics of Complex Systems* (Addison-Wesley, 1997).
[9] S. Boccaletti, G. Bianconi, R. Criado, C. I. del Genio, J. Gómez-Garde nes, M. Romance, I. Sendiña-Nadal, Z. Wang, and M. Zanin, Phys. Rep. **544**, 1–122 (2014).
[10] M. De Domenico, *et al.*, Phys. Rev. X **3**(4), 041022 (2013).
[11] M. Kivela, *et al.*, J. Complex Netw. **2**(3), 203–271 (2014).
[12] F. Battiston, V. Nicosia, and V. Latora, Phys. Rev. E **89**(3), 032804 (2014).

[13] R. J. Sánchez García, E. Cozzo, and Y. Moreno, Phys. Rev. E **89**, 052815 (2014).
[14] B. Min, S. Do Yi, K.-M. Lee, and K.-I. Goh, Phys. Rev. E **89**, 042811 (2014).
[15] R. Criado, J. Flores, A. García del Amo, J. Gómez-Gardeñes, and M. Romance, Int. J. Comput. Math. **89**(3), 291 (2012).
[16] R. Criado, M. Romance, and M. Vela-Pérez, Int. J. Bif. Chaos **20**, 877 (2010).
[17] S. Boccaletti, R. Criado, J. Pello, M. Romance, and M. Vela-Pérez, Int. J. Bif. Chaos **19**(02), 727 (2009).
[18] J. F. Donges, H. C. H. Schultz, N. Marwan, Y. Zou, and J. Kurths, Eur. Phys. J. B **84**, 635 (2011).
[19] J. Gómez-Garde nes, M. Romance, R. Criado, D. Vilone, and A. Sánchez, Chaos **21**, 016113 (2011).
[20] J. Gómez-Garde nes, I. Reinares, A. Arenas, and L. M. Floría, Nature Sci. Rep. **2**, 620 (2012).
[21] E. A. Leicht and R. D'Souza, e-print arXiv:0907.0894.
[22] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin, Nature **464**, 1025 (2010).
[23] C. D. Brummitt, K.-M. Lee, and K.-I. Goh, Phys. Rev. E **85**, 045102(R) (2012).
[24] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J. P. Onnela, Science **328**, 876 (2010).
[25] R. Criado, J. Flores, A. García del Amo, M. Romance, E. Barrena, and J. A. Mesa, Chaos **26**(6), 065309 (2016).
[26] S. Brin and L. Page, Comput. Netw. **30**, 107 (1998).
[27] S. Brin, L. Page, R. Motwani, and T. Winograd, *The PageRank citation ranking: Bringing order to the web*, Technical Report, Standford InfoLab, 1999.
[28] P. Boldi, M. Santini, and S. Vigna, ACM Trans. Inf. Syst. **27**(4), 19:1 (2009).
[29] M. De Domenico, A. Solé-Ribalta, E. Omodei, S. Gómez, and A. Arenas, Nat. Commun. **6**, 6868 (2015).
[30] S. Fortunato, M. Bogu ná, A. Flammini, and F. Menczer, Lect. Notes Comput. Sci. **4936**, 59–71 (2008).
[31] E. García, F. Pedroche, and M. Romance, Linear Algebra Appl. **439**, 640 (2013).
[32] F. Pedroche, E. Garcia, M. Romance, and R. Criado, J. Comput. Appl. Math. **330**, 1029–1039 (2018).
[33] T. H. Haveliwala, S. Kamvar, and G. Jeh, Technical Report, Stanford University, 2003.
[34] V. Chapela, R. Criado, S. Moral, and M. Romance, *Intentional Risk Management through Complex Networks Analysis* (Springer Briefs in Optimization, 2015).
[35] S. Moral, V. Chapela, R. Criado, A. Pérez, and M. Romance, Netw. Heterog. Media **10**(1), 195–208 (2015).
[36] F. Pedroche, M. Romance, and R. Criado, Chaos **26**(6), 06530 (2016).
[37] A. Halu, R. J. Mondragón, P. Panzarasa, and G. Bianconi, PLoS One **8**(10), e78293 (2013).
[38] M. De Domenico, A. Solé-Ribalta, S. Gómez, and A. Arenas, PNAS **111**(23), 8351–8356 (2013).
[39] G. M. Del Corso and F. Romani, Inf. Sci. **329**, 619–637 (2016).
[40] A. Solé-Ribalta, M. De Domenico, S. Gómez, and A. Arenas, Physica D **323–324**, 73–79 (2016).
[41] A. Solé-Ribalta, M. De Domenico, S. Gómez, and A. Arenas, in *Proceedings of the 2014 ACM Web Science Conference* (2014), pp. 149–155.
[42] A. N. Langville, C. D. Meyer, *Google's PageRank, Beyond: The Science of Search Engine Ranks* (Princeton University Press, 2006).
[43] S. H. Whitney, Am. J. Math. **54**(1), 150–168 (1932).
[44] M. Aigner, Math. Z. **102**(1), 56–61 (1967).
[45] J. Bagga, Int. J. Math. Math. Sci. **29**, 1509–1521 (2004).
[46] E. Estrada, N. Guevara, and I. Gutman, J. Chem. Inf. Comput. Sci. **38**(3), 428–431, (1998).
[47] R. L. Hemminger and L. W. Beineke, *Line graphs and line digraphs, Selected Topics in Graph Theory*, edited by W. B. Lowell and R. J. Wilson, (Academic Press, New York, 1978), pp. 271–305.
[48] A. J. Hoffman and D. K. Ray-Chaudhuri, Trans. Amer. Math. Soc. **116**, 238–252 (1965).
[49] P. G. Lehot, J. ACM. **21**(4), 569–575 (1974).
[50] M.-B. Hua, R. Jianga, R. Wang, and Q.-S. Wu, Phys. Lett. A **373**, 2007–2011 (2009).
[51] P. Crucitti, V. Latora, and S. Porta, Chaos **16**, 015113 (2006).
[52] P. Crucitti, V. Latora, and S. Porta, Phys. Rev. E **73**, 036125 (2006).
[53] P. Crucitti, V. Latora, and S. Porta, Physica A **369**, 0411241 (2006).
[54] S. Porta, P. Crucitti, and V. Latora, Environ. Plann. B Plann. Des. **33**(5), 705–725 (2006).
[55] T. S. Evans and R. Lambiotte, Phys. Rev. E **80**, 016105 (2009).
[56] T. S. Evans and R. Lambiotte, Eur. Phys. J. B **77**, 265–272 (2010).
[57] B. Hillier, Enrviron. Plann. B **26**, 169 (1999).
[58] J. Nacher, N. Ueda, T. Yamada, M. Kanehisa, and T. Akutsu, BMC Bioinformatics **5**, 207 (2004).
[59] J. C. Nacher, T. Yamada, S. Goto, M. Kanehisa, and T. Akutsu, Physica A **349**, 349 (2005).
[60] A. Manka-Krason, A. Mwijage, and K. Kulakowski, Comput. Phys. Commun. **181**, 118–121 (2010).
[61] J. Anez, T. De La Barra, and B. Pérez, Trans. Res. B **30**, 209–216 (1996).
[62] D. Volchenkov and Ph. Blanchard, e-print arXiv:0710.5494.
[63] E. Barrena, A. De-Los-Santos, J. A. Mesa, and F. Perea, Eur. Phys. J. Special Topics **215**, 93–108 (2013).
[64] E. Barrena, A. De-Los-Santos, G. Laporte, and J. A. Mesa, Netw. Heterog. Media **10**(1), 1–16 (2015).
[65] S. Derrible and C. Kennedy, Physica A **389**, 3678–3691 (2010).
[66] R. Criado, J. Flores, A. García del Amo, and M. Romance, J. Comput. Appl. Math. **235**(3), 1775–1780 (2011).
[67] R. Criado, J. Flores, A. García del Amo, and M. Romance, Netw. Heterog. Media **7**(3), 373–384 (2012).
[68] R. Criado, M. Romance, and A. Sánchez, Int. J. Bif. Chaos **22**(7), 1250163 (2012).
[69] R. Criado, J. Flores, A. García del Amo, and M. Romance, Int. J. Comput. Math. **91**(2), 304–314 (2014).
[70] V. Latora and M. Marchiori, Phys. Rev. Lett. **87**(19), 198701 (2001).
[71] F. Pedroche, "Métodos de cálculo del vector PageRank," Bol. Soc. Esp. Mat. Apl. **39**, 7–30 (2007).
[72] Metro de Madrid official website https://www.metromadrid.es/.
[73] http://www.mic-ro.com/metro/table.html
[74] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra* (SIAM, Philadelphia, 2000).